社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2004－SLP－54 (19)
2004／12／21

SNR and sub-band SNR estimation based on Gaussian mixture modeling in the log power domain with application for speech enhancements

Tran HUY DAT†, Hiroshi FUJIMURA†, Kazuya TAKEDA†, and Fumitada ITAKURA††

† Graduate School of Information Science, Nagoya University Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan

†† Graduate school of Information Engineering, Meijo University Shiogamaguchi , Tempaku-ku, Nagoya, 468-8502, Japan

E-mail: †{dat,fujimura,takeda}@ sp.is.naoya-u.ac.jp, ††itakuraf@ccmfs.meijo-u.ac.jp

Abstract This work presents a flexible blind SNR estimation method based on Gaussian mixture modeling (GMM) in the log-power domain. Considering the local noise and noisy speech powers as two log-normal distributed random variables, their distribution parameters are estimated via the EM algorithm and used to derive the segmental SNR, which is defined as the expectation distance between two subspace distributions. A compensation mode for the estimation under low SNR conditions ais also proposed. The experimental results, evaluated on the AURORA2 database, show the more consistency of proposed estimation method in both the noise conditions, compared to the conventional methods. The second application presented in this work is sub-band SNRs estimations for speech enhancement systems. Here the GMM is applied to each frequency bin, and then two methods of the sub-band SNRs estimation are proposed by using the maximum a posterior (MAP) decomposition and cumulative distribution function (CDF) equalization. Furthermore, the sub-band SNR is used for the Wiener filtering systems. The evaluation experiments demonstrate the improvements of the proposed speech enhancement method in both segmental SNR and ASR performances.

Key words Gaussian mixture modeling, Signal to noise ratio, sub-band SNR, Wiener filtering.

1. Introduction

SNR estimation is an important problem in speech processing. The SNR is frequently used in the speech data collection [1] which is an indispensable task for the development of automatic speech recognition systems and the sub-band SNRs are used in speech enhancement systems [2]. In real conditions, the SNR estimation of noisy speech is extremely difficult because: (1) clean signal and noise references are not available, and (2) voice activity reference is unknown. Conventionally, the SNRs estimations are based on separation of the noise and speech subspaces. The separation can be achieved on the speech via the voice activity detection (VAD) techniques, or on basic of a histogram, estimated from a given noisy speech. For the SNR estimation. a popular method is proposed in [3], where the VAD is detected by applying a threshold based soft decision to the power histogram. The noise and speech levels then are calculated by averaging the powers with and without speech activities. The other application [4] proposed the separation of subspaces in the histogram on basic of a heuristic setting a raised cosine function in the lower part of the histogram,according to the noise subspace. A drawback of these methods is that, they require the separation of noise and speech subspaces to calculate the SNR estimations. Therefore, the sufficiency of each method are dependent very much on this separation processes, which work well only under high SNR conditions. Moreover the SNR defined in [3] is in fact the total to noise ratio (TNR) which even theoretically can not approximate the SNR at the low SNR conditions, and the SNR defined in [4] is peak to noise level which is different from the standard definition of the segmental SNR.

In this work we consider the logarithm of local noise and noisy speech power as two random variables and use the natural mixture structure of speech signals (i.e the consist of the pauses) to model and estimate the distribution parameters. The Gaussian mixture modeling (GMM) in the

log power domain with the EM parameter estimation are used in this work. The SNR can be described as an expectation of a nonlinear function of two random variables of local noise and noisy speech powers and can be estimated statistically and without knowledge of VAD. Note that here we estimate the two subspaces of noise and noisy speech as defined via their distribution parameters using the maximum like lihood criteria, but no longer an exact separation of them is required. Furthermore, the SNR can be estimated in two modes. In the easy mode, the SNR is defined as the deference between the means of subspaces distributions. Although this mode is very simple it can be successfully used for the real conditions, when the SNR is greater than approximately 10 dB. In the compensation mode we express the SNR as the expectation of a nonlinear function of local noise and noisy speech powers. The distribution parameters are used in the compensation mode to yield the estimation more correct, particularly under the low SNR conditions.

The second approach, presented in this work is sub-band SNR estimations with applications to the noise suppression. The single microphone noise suppression technique requires the sub-band SNR estimations. For the sub-band SNR estimations, various approaches have been proposed without explicit knowing of VAD [2],[5],[6],[7] and both of them are performed in along with a deterministic manner with an assumption of stationary or relative stationary of noise in the sub-bands.

As same as in the SNR estimation, the main point of our proposed method is using the long time statistic of the speech and noise in sub-bands which can be estimated by using the natural mixture structure of the speech (i.e. always consist the pause durations). Furthermore, two statistical estimation methods for sub-band SNRs based on the maximum a posterior (MAP) decomposition and cumulative distribution function equalization are investigated and the Wiener filtering speech enhancements are implemented.

The organization of this paper is as follows. In section 2 we discuss the stochastic view of SNR definitions and Gaussian mixture modeling (GMM) in the log-power domain, and present the estimation of the segmental SNR and report the experimental results, compared the proposed and conventional methods. In section 3 we apply the GMM in each frequency bin ans presents two methods for the sub-band SNRs estimations and apply them to the Wiener filtering. The effectiveness of the proposed methods are investigated using the standard AURORA 2. Section 4 summarizes the work.

2. SNR ESTIMATIONS

2.1 Stochastic view of SNR definitions

Several measurements of SNR are used as the speech quality indexes. Since a speech signal is short-time stationary, the segmental SNR is advantageously used [1].

Originally, the segmental SNR is calculated in speech active frames and is noted by

$$SNR_{seg} = \frac{1}{L}\sum_{i=1}^{L} 10\log_{10}\frac{P_S(i)}{P_N(i)}, \tag{1}$$

where $P_S(i)$ and $P_N(i)$ are respectively the clean reference speech and the noise power at the i-th speech active frame, and thus are called frame powers. When clean speech is not available, it is Moree suitable to use the total (noisy speech) to noise ratio (TNR), and it is denoted by:

$$NRTNR_{sage} = \frac{1}{L}\sum_{i=1}^{L} 10\log_{10}\frac{P_X(i)}{P_N(i)}, \tag{2}$$

where $P_X(i)$ is the frame power in the i-th speech active frame. Due to the independence of speech and noise we assume that:

$$P_X(i) = P_S(i) + P_N(i). \tag{3}$$

Under a high S condition (more than 10dB), the SNR is well approximated by the TNR, and in some applications, the TNR is defined as the SNR [2].

When the frame number L is large, the averaging in (1) will converge to the expectation. The segmental SNR can be denoted in a term of expectations:

$$SNR_{seg} = \left\langle 10\log_{10}\frac{P_S}{P_N}\right\rangle = \langle 10\log_{10}P_S\rangle - \langle 10\log_{10}P_N\rangle, \tag{4}$$

Analogously, the TNR can be denoted in the stochastic form as follows:

$$TNR_{seg} = \left\langle 10\log_{10}\frac{P_X}{P_N}\right\rangle = \langle 10\log_{10}P_X\rangle - \langle 10\log_{10}P_N\rangle, \tag{5}$$

On the basic of (4-5), the TNR and SNR measurements are the expressions of the expectations of the noise and the noisy speech (or speech) powers and therefore are defined if their distributions are known or being estimated.

2.2 Gaussian mixture modeling

The basic concept of our approaches that if we consider the logarithm of noisy speech and the noise frame powers as two random variables (Figure 1), their distribution parameters can be estimated on the basic of observed noisy speech by using mixture modeling. Note that, the natural

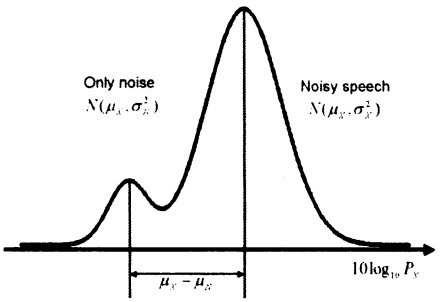

Fig. 1 Two Gaussian mixture modeling on log-power domain.

mixture structure of speech signals is given according to the existing of the pause durations. In this work, we use a two-component Gaussian mixture for which the pdf is denoted as

$$p(x) = \alpha_0 \mathrm{N}\left(x, \mu_N, \sigma_N^2\right) + \alpha_1 \mathrm{N}\left(x, \mu_X, \sigma_X^2\right). \quad (6)$$

Here $\alpha_1 + \alpha_0 = 1$ and $\mathrm{N}\left(x, \mu, \sigma^2\right)$ denotes the Gaussian distribution.

The logarithm domain can be considered as a compressed operator, which reduce the dynamic range of speech and provide more stationary of observations. It has been proved that the EM algorithm [8] produces monotonically non-decreasing sentences of log-likelihood. Therefore from an initial set of parameters, the algorithm converges to the local maximum of the log-likelihood. The initial parameters are chosen by the standard K-means method [9]. For our procedure, we verified the EM convergence after 5-7 iterations.

2.3 Easy mode of the segmental SNR estimation

On the basis of the definition given in (4), the segmental TNR is considered equal to the difference between the two estimated means,

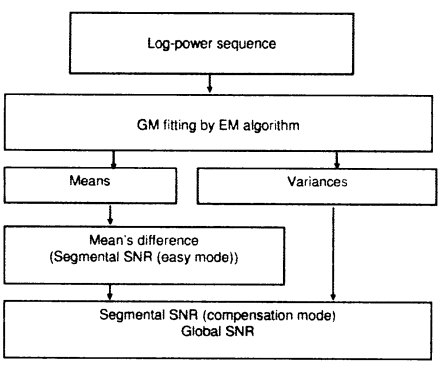

Fig. 2 Block diagram processing of SNR estimation based on GMM on log-power domain

$$TNR_{seg} = \mu_X - \mu_N, \quad (7)$$

where $\mu_X > \mu_N$.

Note that, under the high SNR condition $logP_X \approx \log P_S$, therefore the segmental TNR approximates the SNR and can thus be used a a segmental SNR estimation.

In this work we use (7) for the easy mode of segmental SNR estimation. Note that, although this mode is simple, it works well under the SNR condition higher than 10dB, and therefore can be applied in lot of real conditions.

2.4 Compensation mode of the segmental SNR estimation at low SNR conditions

Under the low SNR conditions, the logarithm of noisy speech power cannot be use to approximate the logarithm of clean speech power, therefore the easy mode yield a significant errors in these cases. Therefore, to archive an accurate estimation, we implement a compensation mode of segmental SNR estimation as below.

Assuming the noisy speech power is a superposition of the clean speech and noise powers in each frame and substituting (3) into (4), the segmental SNR is denoted by

$$SNR_{seg} = \left\langle 10\log_{10}\left(\frac{P_X}{P_N} - 1\right)\right\rangle. \quad (8)$$

Furthermore, the estimated distribution parameters will be used to calculate the nonlinear moment in (8). Recalling the Gaussian mixture model, the two random variables in (6) have Gaussian distributions:

$$\begin{aligned} 10\log_{10} P_N &\sim \mathrm{N}\left(x, \mu_N, \sigma_N^2\right), \\ 10\log_{10} P_X &\sim \mathrm{N}\left(x, \mu_X, \sigma_X^2\right), \end{aligned} \quad (9)$$

and therefore, their difference is also Gaussian distributed:

$$10\log_{10}\frac{P_X}{P_N} \sim \mathrm{N}\left(x, \mu_X - \mu_N, \sigma_X^2 - \sigma_N^2\right). \quad (10)$$

The expectation of the nonlinear function (8) has no closed form but can be approximated using an asymptotic expansion:

$$\ln\left(e^r - 1\right) \approx r - 0.7e^{-r} - 0.9e^{-2r} - e^{-3r}. \quad (11)$$

Note that the approximation error is less than 1% at$r >$ 0.12, i.e. $10\log_{10}\frac{P_S}{P_N} > -9.78$dB. Since in this work our interest is the signals with SNR from 0 to 20 dB, the error of approximation (11) is negligible. The expectation of the approximation (11) when r is a Gaussian random variable can easily be calculated. A closed form of the segmental SNR estimation is derived as follows:

$$SNR_{seg} = \frac{10}{\ln 10}\left\{\begin{array}{l} m - 0.7\exp\left[-\left(m - \frac{d}{2}\right)\right] \\ -0.9\exp\left[-2\left(m - d\right)\right] \\ -\exp\left[-3\left(m - \frac{3d}{2}\right)\right] \end{array}\right\}, \quad (12)$$

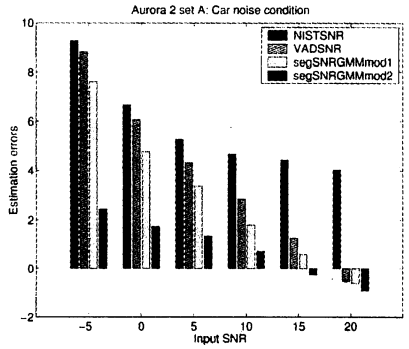

Fig. 3 Estimation errors of the segmental SNR estimations in the car noise condition

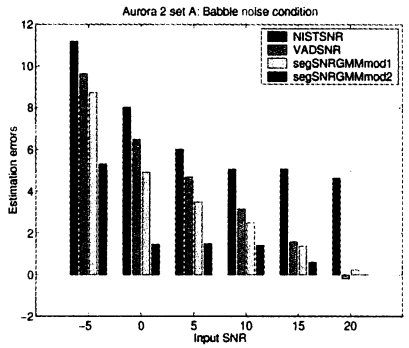

Fig. 4 Estimation errors of the segmental SNR estimations in the babble noise condition

where

$$m = \frac{\ln 10}{10}\left(\mu_X - \mu_N\right), \quad d = \left(\frac{\ln 10}{10}\right)^2 \left(\sigma_X^2 - \sigma_N^2\right). \tag{13}$$

Equation (12) is used in our compensation mode for the segmental SNR estimation. Obviously, the first term in (12) coincides to the segmental TNR (easy mode estimation), and the consequent terms are negligible when m is large but become considerable at small value of m according to the low TNR conditions.

2.5 Experiment

The block diagram of the processing procedure is shown in Figure 2. Note that the hamming window with the frame length of 4ms and a frame shift of 2ms is used to provide the enough large sample number, which is necessary for the EM algorithm. With this frame length, the minimum required length of signal is verified from the experiments to approximately 1 second at 8kHz sampling frequency.

For the evaluation of segmental SNR estimation, the standard AURORA2 data [10] is used. We test the proposed algorithm in both easy and compensation mode

SNR	NISTSNR	VADSNR	segSNRmod1	segSNRmod2
-5dB	11.2±3.1	10.1±2.8	9.2±2.1	5.1±1.9
0dB	7.8±3.0	6.5±1.9	5.1±1.5	2.1±1.7
5dB	5.8±2.8	4.9±1.5	3.8±1.6	1.5±1.0
10dB	4.9±2.1	3.0±1.7	2.2±0.9	1.1±0.9
15dB	4.3±2.2	1.2±0.5	0.9±0.4	0.4±0.6
20dB	4.0±1.9	-0.3±0.5	-0.2±0.2	-0.4±0.3

Tab. 1 Average SNR estimation evalution on AURORA2

(segSNRGMMmod1 and segSNRGMMmod2). For the reference, the segmental SNR based on VAD (VADSNR) using a threshold decision [3] and the NISTSNR [4] are also implemented. Table 1 shows the average results obtained in each the noise conditions and the Figure 3-4 show examples for the car noise and babble noise conditions. The easy mode of segSNRGMM (segmental TNR) yields more accurate estimation than conventional methods. However, the estimation error of the easy mode estimation (segmental TNR) is quite significant at low SNR conditions. The estimation errors of the compensation mode of the segSNRGMM are less than 1.5 dB when the original SNR is greater than 5dB. At the high SNR conditions (20dB), the segSNRGMMs are over estimated. This can be explained by that, the speech pause subspace becomes comparable to the noise therefore the estimated SNR are over comparing to the "true", which does not take into account the noise of speech pause durations.

The NISTSNR over estimated the segmental SNR in both cases. Note that, because this measurements is defined as the peak speech level to the average noise level ratio, the bias of NISTSNR to segmental SNR depends on the variance of speech distribution which are deter minded from each speech data.

One important thing is the flexibility and relative robustness of the segmental SNRGMM estimation, which can be carried out in real speech data, without any references, and without any use of the control parameter.

3. Sub-and SNR estimation based noise suppression

Another approach is sub-band SNR estimations for the noise suppression filtering. Conventional Wiener filtering on frequency domain is optimal filtering for the Gaussian model of speech and noise [2], where the noise suppression gain function is defined via the sub-band SN Rs as follows:

$$G(l,k) = \frac{P_S(l,k)}{P_X(l,k)} = 1 - \frac{1}{\frac{P_X(l,k)}{P_N(l,k)}} = 1 - \frac{1}{TNR(l,k)}, \tag{14}$$

where TNR is the sub-band TNR in the linear domain and is defined at frame index l on frequency bin k. Note that since there is no need here to differ the SNR and TNR terminology, the sub-band SNR estimation is understood as the estimation of TNR. Unlike in the conventional methods, in this work we develop the statistical estimation methods for the sub-band SN Rs, using the estimated distribution streamers. Denote the Gaussian distributions of the GMM in the log-sub-band power domain as follows

$$\begin{aligned} \ln P_N(k) &\sim \mathrm{N}\left(x, \mu_N(k), \sigma_N^2(k)\right), \\ \ln P_X(k) &\sim \mathrm{N}\left(x, \mu_X(k), \sigma_X^2(k)\right). \end{aligned} \tag{15}$$

As in the SNR estimation the distribution parameters in (15) are estimated via the EM algorithm. Furthermore,we investigate two approach of the maximum posterior estimation and the cumulative distribution function mapping methods for the sub-band SNRs estimation.

3.1 MAP estimation of sub-band TNR

Express the noisy speech log-power in terms of log-noise power and TNR as follows:

$$\ln P_X(l,k) = \ln P_N(l,k) + \ln TNR(l,k), \tag{16}$$

where $TNR(l,k)$ denote the TNR in the linear domain.

The MAP estimation of logarithm of TNR is denoted by

$$\widehat{\ln TNR} = \max_{\ln TNR} p(\ln P_X | \ln TNR) p(\ln TNR). \tag{17}$$

As in the SNR estimation, the TNR (in this case logarithm of linear TNR) is considered as a Gaussian random variable

$$\ln TNR(l,k) \sim N\left(\mu_X - \mu_N, \sigma_X^2 - \sigma_N^2\right). \tag{18}$$

Substituting (15) and (16) into (17) yields the estimation of the sub-band TNR as follows:

$$\widehat{TNR(l,k)} = \exp\left[\begin{array}{l} \mu_X(k) - \mu_N(k) + \\ \frac{\sigma_X^2(k) - \sigma_N^2(k)}{\sigma_X^2(k)} \left(\ln P_X(l,k) - \mu_X(k)\right) \end{array}\right]. \tag{19}$$

From (19), the sub-band TNR estimation is a non-linear function of the local noisy speech powers and the distribution parameters.

3.2 Cumulative distribution function mapping

Another statistical estimation method can be applied for the sub-band TNR estimation is cumulative distribution function mapping (or equalization) (CDFM). In this estimation scheme, we look for a sub-band TNR estimation in (16) in each frequency bin as a nonlinear function of the logarithm of noisy speech powers

$$\widehat{\ln TNR(l,k)} = g_k(\ln P_X(l,k)). \tag{20}$$

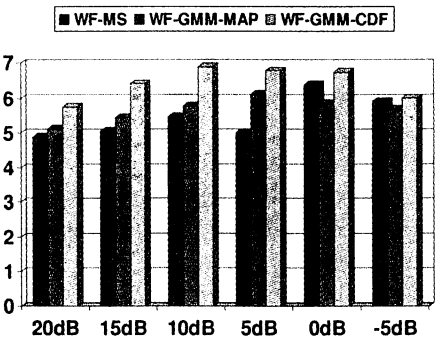

Fig. 5 Segmental SNR improvement comparison of Wiener filtering methods

The criteria to estimation here is that the CDF of the estimated sub-band TNRs must equal to their prior CDF

$$F_{g(x)}(g(x)) = F_s(s). \tag{21}$$

The key point of CDFM method is that the CDF of arbitrary function of a variable is equal to the original CDF

$$F_{g(x)}(g(x)) = F_x(x). \tag{22}$$

From (21) and (22) the estimation is given by mapping (equalization) the prior CDF.

$$g(x) = F_s^{-1}(F_x(x)). \tag{23}$$

The CDFM in (23) is given for the sub-band TNR, where the CDFs are CDF of Gaussian distribution in (15).

3.3 Experiments

According the random behavior of the noise, we never can estimated a exact sub-band TNRs which are the realization of random variables. Therefore the comparison of estimated sub-band TNR and original (if it is available) is not meaning full. However this estimation makes sense in the evaluation performances of total enhancement systems. We implement the Wiener filtering using the sub-band SNRs estimation derived above. The processing diagram is follows. The noisy speech program (spectral magnitude square) is smoothed by the moving average [2] and is divided into sub-bands corresponding to each frequency index k. The GMM fitting via the EM algorithm is applied to each sub-band for the narrow band powers. The sub-band SNRs are estimated according (19) and (23) and furthermore are used for the gain function (14) of the Wiener filtering. For reference, the conventional Wiener filtering based on the minimum statistic noise estimation method is also implemented [7]. The gain function in these cases is derived analogously using (14).

The standard Aurora2 speech data set is used for evaluation. The data has moderate SNR conditions from -5dB to

Training mode	Set A	Set B	Set C	Overall
Multi condition	11.44	12.01	14.68	12.32
Clean	26.84	27.16	17.97	25.20
Average	19.14	19.59	16.33	18.76

Tab. 2 WER evaluation on AURORA2 for the minimum statistic noise estimation based Wiener filtering (WF-MS)

Training mode	Set A	Set B	Set C	Overall
Multi condition	10.08	11.29	13.14	11.18
Clean	25.35	25.77	16.64	23.78
Average	17.72	19.03	14.89	17.48

Tab. 3 WER evaluation on AURORA2 for the Gaussian mixture modeling in sub-band and MAP estimation based Wiener filtering (WF-GMM-MAP)

Training mode	Set A	Set B	Set C	Overall
Multi condition	10.88	10.64	12.04	11.02
Clean	23.68	21.74	15.87	21.34
Average	17.28	16.19	13.96	16.18

Tab. 4 WER evaluation on AURORA2 for the Gaussian mixture modeling in sub-band and CDF mapping based Wiener filtering (WF-GMM-CDF)

20dB in different noise conditions. The two measurements of the segmental SNR and the speech recognition rate are evaluated. Speech recognition experiments are performed on the Aurora 2 connected to the digit recognition task [10]. The digit HMMs are the standard complex back-end models of 16 states , and each state has a 20 components Gaussian mixture with diagonal covariance matrix. The training process is carried out at each front-end before training. The feature vector has 39 components of 12 mean normalization MFCC coefficients together with C0, their first and second derivatives.

Figures 5 shows the overall results of the average segmental SNR improvements over moderate SNR for each noise conditions. From the Figure, the WF-GMM-CDF system performed best, where the relative improvement is up to 2dB compared to the WF-MS method. The SNR improvement of the WF-GMM-MAP method is comparable to the WF-MS at low SNR conditions and little better at the high SNR conditions. The enhanced speech signals are subject to the information listening tests. In general, the musical like artifacts audible level is less at the WF-GMM-CDF but is highest at the WF-GMM-MAP. The results of speech recognition are shown in Table 2-4 for the word error rate. Both the proposed WF-GMM-CDF and WF-GMM-MAP are performed better than the conventional method in clean and multi-condition training.

4. Conclusions

In this study we investigate the Gaussian mixtures modeling of speech on the log-power domain for the SNR estimations. The main point here is stochastic view of local noise and speech power on both broadband and sub-band measures and can be defined via the distribution parameters. No exact VAD is necessary to separate these two subspaces.

The two-components Gaussian mixture modeling is shown to be flexible and effective to estimate the distribution parameters of noise and noisy speech speech subspaces and it can be done on-line from given noisy speech. The proposed segmental SNR estimation method is shown to be more accurate and flexible than VAD based method.

This model can also be successfully applied for the sub-bands SNR estimation with the further application in speech enhancements systems. The experiments show at least the comparable of proposed method in the segmental SNR and ASR measurements and less level of musical like artifacts, compared to conventional method.

In future work we would like to investigate proposed estimation for other feature domain, for example MFCC.

References

[1] J. Hansen, and B. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms", *Proc. IEEE ICSPL*,1998, pp.2819-2822

[2] Y. Ephraim, and D. Malah, "Speech enhancement using a MMSE log-spectral amplitude estimations," *IEEE Trans. ASSP*, Vol. 33, No. 2, pp.443-445, 1985.

[3] A. Korthauer,"Robust estimation of the SNR of noisy speech for the quality evaluation of speech databases," *Proc. IEEE RMSRAC*,1999.

[4] NIST Speech Quality Assurance (SPQA) Package *http://www.nist.gov/speech/tools/index.htm.*

[5] H. Hirsch, and C.Ehrlicher, "Noise estimation techniques for robust speech recognition" *Proc. IEEE RMSRAC*,1999.

[6] R. Martin, "Noise power spectral estimation based on optimal smoothing and minimum statistics," *IEEE Trans. ASSP*, Vol. 9, No.5, pp.504-512, 2001.

[7] I. Cohen "Noise spectrum estimation in adverse environments," *IEEE Trans. ASSP*, Vol 20, 2002.

[8] A. Dempster, N. Laird, and D. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm", *J.Royal Statist.Soc.*, Ser.B , Vol.39, 1977

[9] S. Dasgusta, "Learning Gaussian mixtures," *Proc. IEEE ICASSP*,1995,pp.153-156

[10] H. Hirsch, D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *in ISCA ITRW ASR*, 2000.